The IJB-A Face Verification Challenge
Performance Report

Patrick Grother and Mei Ngan

Caution: This report quantifies face recognition performance using data supplied by external research and development organizations. Its results are derived from self-administered experiments on the fully public IJB-A dataset. As such the results can be manipulated by various means that may not be operationally realistic. Therefore, end users of face recognition technology should prefer results from NIST's ongoing sequestered testing campaigns, FRVT or FIVE, or on similar independent evaluations of face recognition. Developers whose algorithms exhibit good performance here are encouraged to submit their algorithms into those sequestered tests programs.

This report is generated automatically. It will be updated as new algorithms are evaluated, and as new analyses are added. Automated notifications can be obtained via the mailing list. Correspondence should be directed to the authors via FaceChallenges@nist.gov.

This report was last updated on July 23, 2015.

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

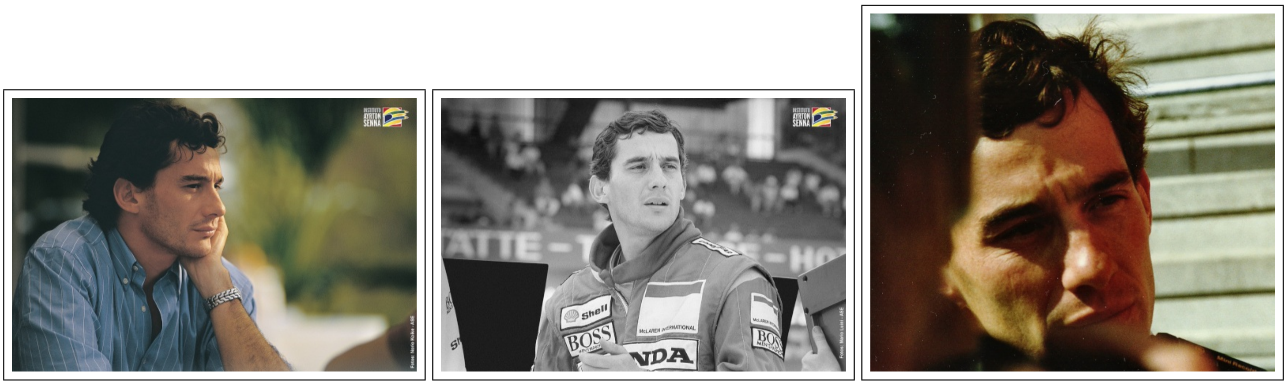

Figure 1: Three images of one subject in the IJB-A dataset. The entire dataset is available online. Many photos were taken by photo journalists and, as such, are well exposed, well focused, and deemed suitable for public display. For face recognition, they nevertheless remain challenging due to wide variations in pose, illumination, expression and occlusion.

1 Introduction

Three IARPA Janus Benchmark A challenges are described by Klare et al. in the paper *Pushing the Frontiers of Unconstrained Face Detection and Recognition*[2]. The first of these, the IJB-A 1:1 challenge, quantifies performance of face verification algorithms ("same person or not?") on challenging photo-journalism images of the kind shown in Figure 1. They are considerably more difficult to recognize than the portraits mandated by facial recognition standards[1].

IJB-A 1:1 is a "take-home" test in that it is based on fully public data. It follows the design of the LFW protocol in requiring many pairs of samples to be compared in isolation[2]. This corresponds to recognition tasks like passport verification or forensic comparison where there is just a pair of samples and no central database or gallery.

The IJB-A 1:1 challenge departs from LFW as follows:

- **Face selection**: LFW contains faces that could be detected with the Viola-Jones face detection algorithm. This limits difficulty. IJB-A on the other hand, uses manually located and annotated faces.
- **Landmarks**: The IJB-A tests include landmark coordinates (eyes and nose) whereas LFW provides just raw images, and aligned (funneled) images.
- **Multi-image samples**: LFW compared single images. IJB-A uses richer samples containing $1 \leq K \leq 202$ images, including frames from video sequences.
- **More impostor pairs**: IJB-A 1:1 uses many more impostor comparisons that genuines. In LFW, the ratio was 1 which precluded computation of false match rates at usefully low values.

2 Metrics

This section describes the one-to-one verification accuracy metrics present in this report.

[1] NIST maintains a challenge for such images based on the mugshots of NIST Special Database 32 ("MEDS")[1]. This is intended as a stepping stone prior for developers prior to entering NIST's ongoing fully sequestered FRVT verification test.

[2] IJB-A 1:1 does not cross-compare galleries and probesets; it has no concept of such. It does not attempt to measure both verification and identification accuracy from the same similarity score matrix; it does not pin the prior probabilities of impostor vs. genuine pairs i.e. O(n^2) vs. O(n).

2.1 Quantifying false acceptance

False acceptance is computed over N comparisons. Each comparison involves a pair of samples, one from each of two different individuals. Each comparison yields a single non-negative scalar similarity score. The false match rate (FMR) is defined as the proportion of scores that are at or above threshold T.

$$\text{FMR}(T) = 1 - \frac{1}{N} \sum_{i=1}^{N} H(s_i - T) \tag{1}$$

where H is the Heaviside unit step, and s_i is the score from the i-th impostor comparison.

2.2 Quantifying false rejection

False rejection is computed over M comparisons. Each comparison involves a pair of samples. Both samples come from a single individual. Each comparison yields a single non-negative scalar similarity score. The false non-match rate is defined as the proportion of scores that are below a threshold T.

$$\text{FNMR}(T) = \frac{1}{M} \sum_{i=1}^{M} H(s_i - T) \tag{2}$$

with s_i being the score from the i-th genuine comparison.

The common term true accept rate (TAR) is a synonym for the complement 1-FNMR. This document uses neither TAR nor false accept rate (FAR), as these are reserved[3] for transactional error rates involving potentially several attempts to use a biometric system. The terms here, FNMR and FMR, are *matching* error rates, befitting IJB-A.

2.3 Quantifying failure to enrol

The above definitions assume that each comparison produces a score. Indeed the IJB-A protocol requires a complete set of scores to be submitted. This is operationally atypical because: a) some algorithms electively refuse to enrol imagery that is too poor to process; b) face or landmark detection can fail; c) feature extraction fails; and d) software throws an exception. Together these outcomes are quantified by the failure-to-enrol rate (FTE), a term whose definition is overloaded in the literature, and in practice.

This report includes two statements of the failure to enrol rate (FTE).

- ▷ **Empty templates**: The proportion of all templates that are empty, where empty is defined as having size below 32 bytes. This value is used heuristically to handle implementations that sometimes fail to produce a viable template but nevertheless include a short header.
- ▷ **Failed comparisons**: The proportion of comparisons that give a non-zero return code, or a negative similarity score.

The consequences of FTE operationally should depend on the application. A one-to-one access control system would reject the presentation, and allow limited re-submissions of new face samples. In negative identification systems, where subjects make an implicit claim not to be present in a database (e.g. deportees), the correct action would be to investigate the sample for evidence of evasion or tampering. To effect fair comparison of algorithms, failed comparisons must be accounted for. This is done here by setting scores to zero to simulate rejection. This corresponds operationally to a (fortuitously) correct rejection of an impostor, and an incorrect rejection of a genuine user. Note that in the case of negative identification, scores should be set to high values to simulate the triggering of a human investigation. This would correctly flag enroled subjects, and incorrectly flag the non-enroled. Most research reports set scores to zero, as is done here.

This issue presents a gaming opportunity: An algorithm developer can code a strategy for handling low quality images if he knows or assumes the number of impostor comparisons in a test will be larger (or smaller) than the number of genuines. In such cases, his implementation would speculatively and preferentially guess a low (high) score in cases where feature extraction fails, or when image quality is poor. While the IJB-A challenge is wide open to such gaming, there are several mitigations to such strategies, particularly in sequestered tests. Operationally, an off-the-shelf implementation will not know the prior probability of an impostor.

2.4 Score-range normalization

Some graphs in this report include similarity scores normalized via the probability integral transform. If the empirical cumulative distribution of the comparison scores is F, then the new variable

$$s^{\dagger} = F(s) \tag{3}$$

will be uniformly distributed on [0,1] (absent tied scores). This affords comparison between algorithms. If the transform is applied to just the genuine scores, then for a threshold, T, the quantity F(T) is the false non-match rate, FNMR(T), i.e. the fraction of genuine scores below threshold. The function F can be applied also to impostor scores.

3 Results

3.1 Core accuracy statements

The graphs that follow include results for COTS algorithms. The implied comparison with the other algorithms is caveated by the following

- **No training: The COTS algorithms were purely *off-the-shelf* and were not trained in IJB-A data.**
- **Development date: The COTS algorithms were developed 2012-2013. The other algorithms were developed mid 2015.**
- **Landmarks: The COTS algorithms were not able to consume landmark information and did not use it. The COTS algorithms were given cropped faces.**

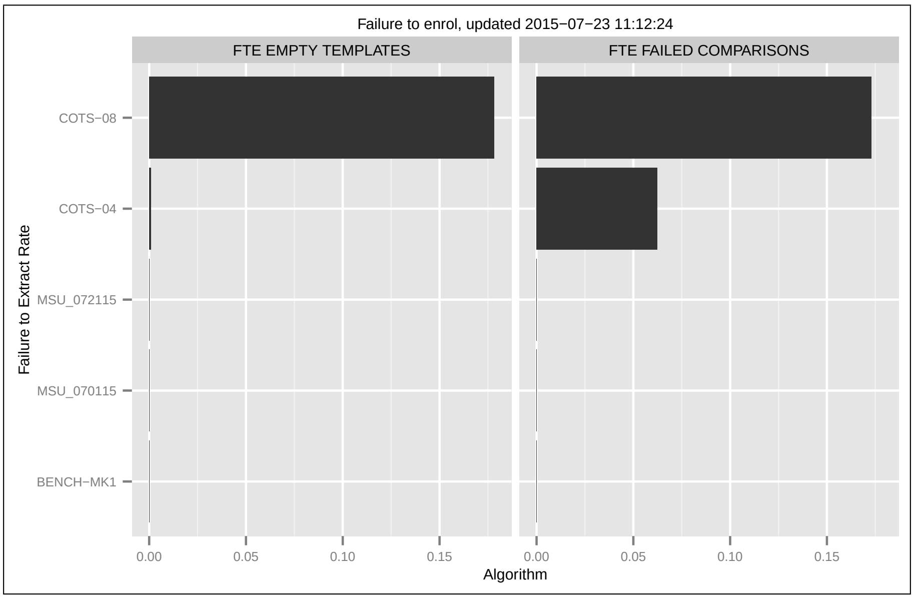

Figure 2: **Failure to enrol:** Per the discussion in section 2.3, the chart quantifies failure to enrol rate (FTE) in two ways: as the proportion of empty templates; and as the proportion of failed comparisons. These two quantities are not independent since empty templates should not yield valid comparison scores.

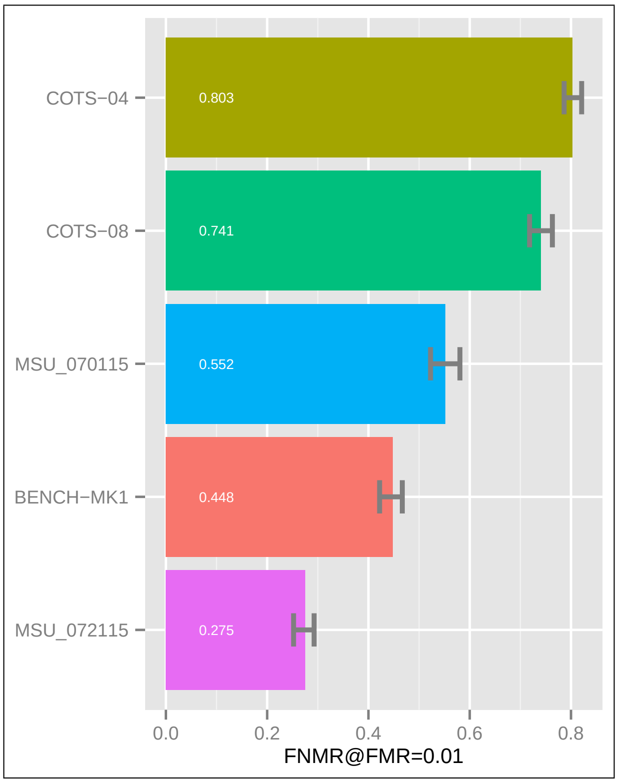

Figure 3: **Leaderboard:** The chart shows the false non-match rate (FNMR) at a false match rate (FMR) of 0.01, as a headline comparative accuracy statement. This FMR value is higher than that typically targeted in operational face verification settings, but is chosen here given the limited number of subjects present in the IJB-A 1:1 dataset. The full error tradeoff characteristics of Figure 4 give accuracy estimates at a broader range of FMR values. **The COTS algorithms were developed before the IJB-A set was developed, and were provided to NIST without any training nor expectation that they would be run on images of this type.**

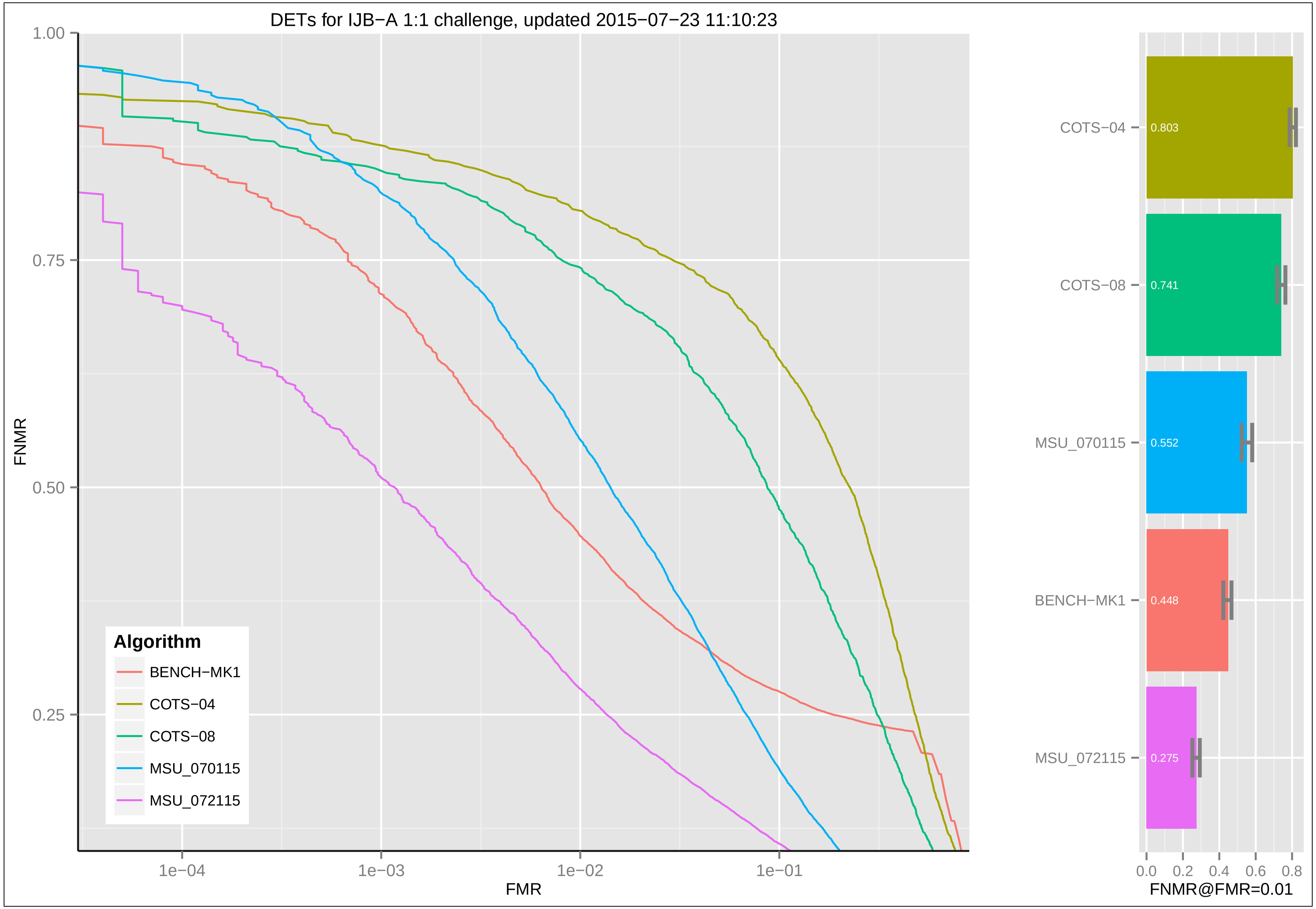

Figure 4: **DETs:** At left, are full error tradeoff characteristics (DETs) for the IJB-A algorithms. At right is a summary statistic corresponding to a vertical slice from the DETs. This is included to ease lookup. While it implies one ranking of the algorithms, it is notable that some DETs cross such that false rejection accuracy depends heavily on the FMR operating point. Algorithm developers can shape the DET characteristics according to a known use-case. Much academic research optimizes at FMR values higher than is operationally useful. **The COTS algorithms did not have access to landmarks. They were developed before the IJB-A set was collected, and were not trained on images of this type.**

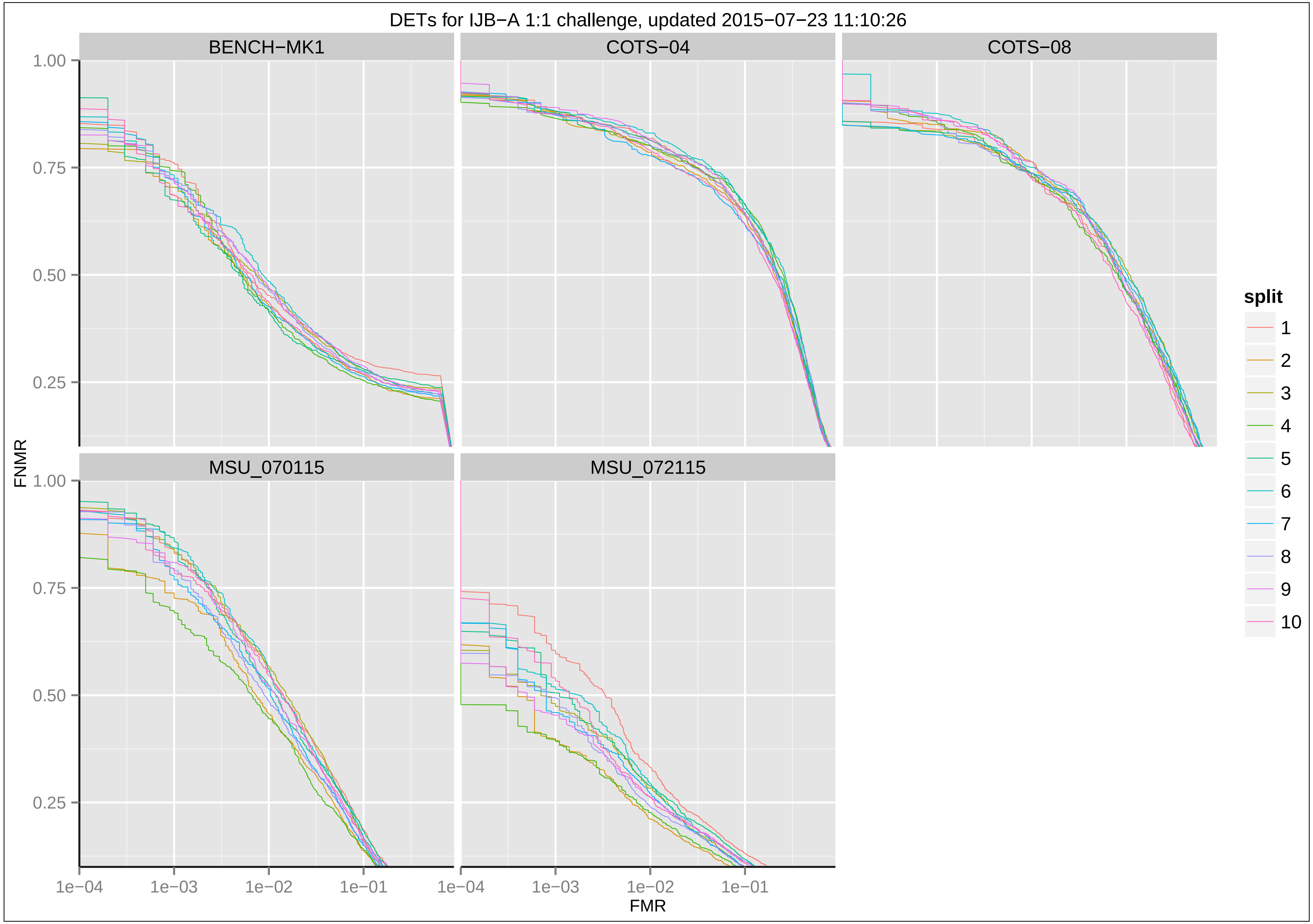

Figure 5: **DETs by split:** DETs, as before, but with individual traces for each of the ten splits present in the IJB-A set.

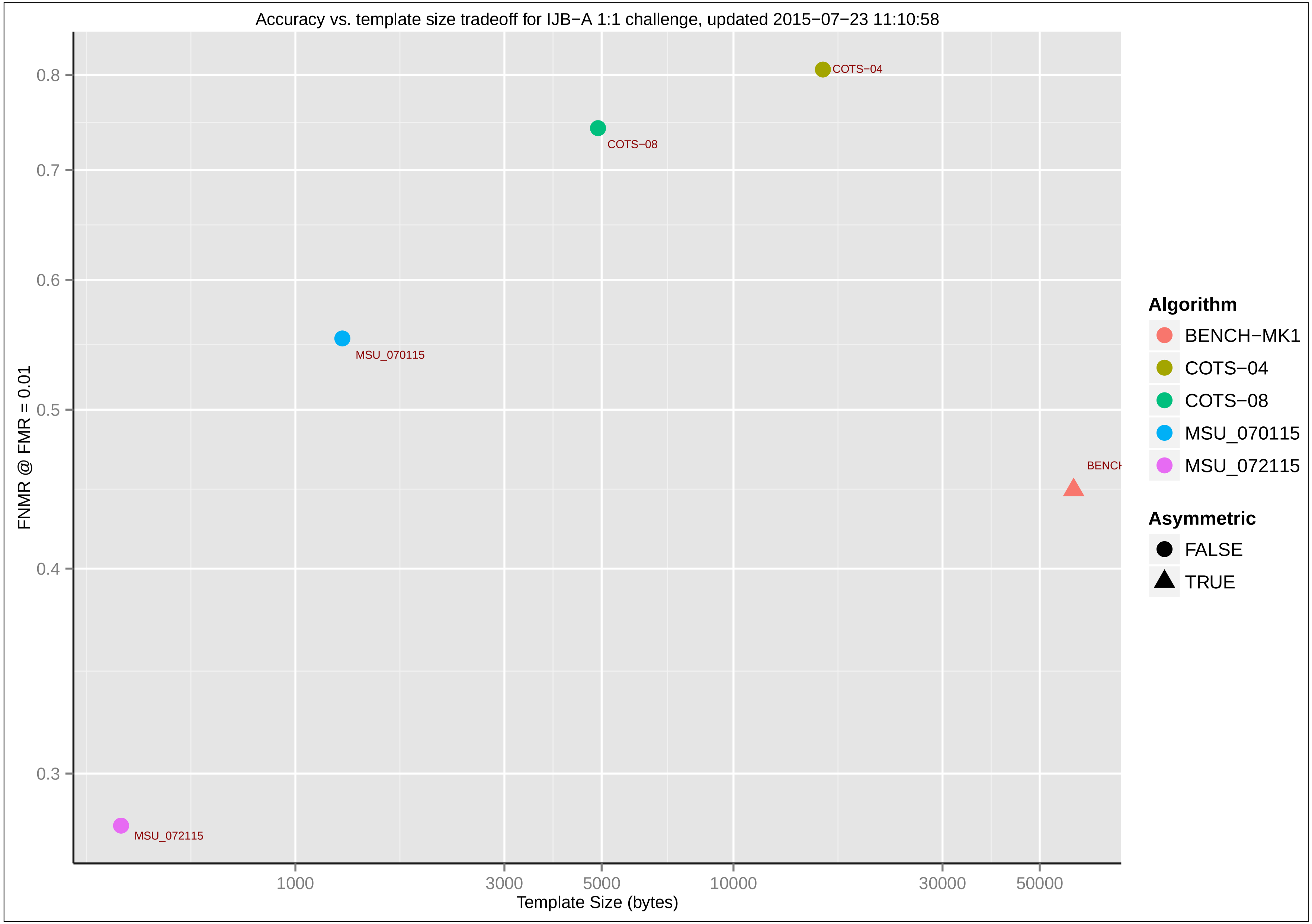

Figure 6: **Accuracy vs. template size tradespace**: The graph plots a summary FNMR value (at FMR = 0.01) against the size of a single-image template. While large templates notionally offer information-rich representations and improved accuracy, they may also suffer curse-of-dimensionality effects. There is no clear industry-wide tradespace evident. An asymmetric algorithm has different enrolment and verification template sizes. **The COTS algorithms did not have access to landmarks. They were developed before the IJB-A set was collected, and were not trained on images of this type.**

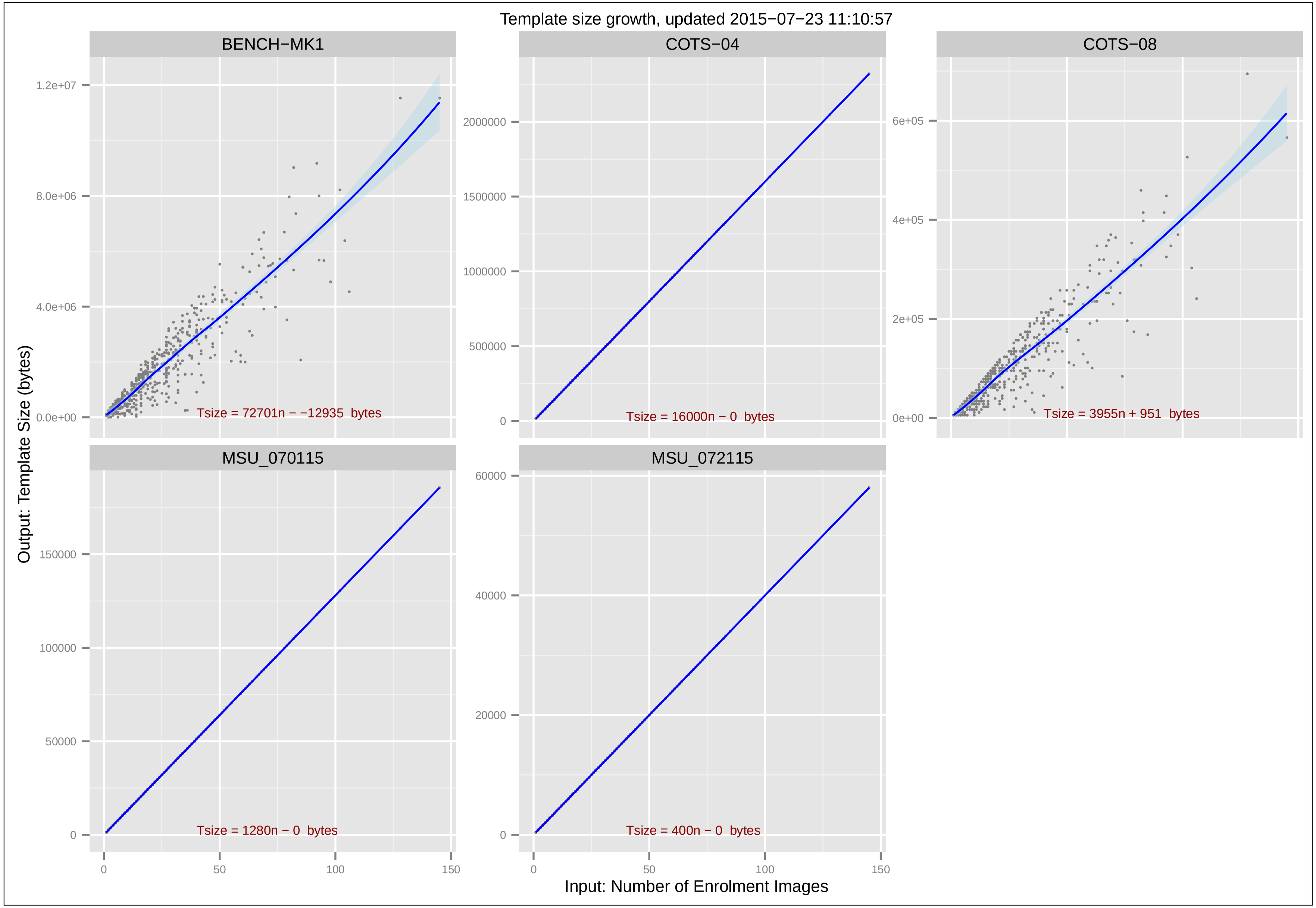

Figure 7: **Template size**: IJB-A compares samples. Each sample is composed of $1 \leq n \leq 202$ faces. Each panel plot shows the growth of the reported total template size as a function of the number of separate images passed to the feature extraction routine. The blue line shows a smoothed dependence. The red text gives an OLS linear growth model; its gradient gives the marginal i.e. template size for one additional image. The model is incorrect in some cases. An O(n) dependence is typical, reflecting extraction of features from each frame independently. An O(1) dependence is consistent with either a) building of a fixed-size model from arbitrary amounts of imagery or b) selection of a single best image for feature extraction.

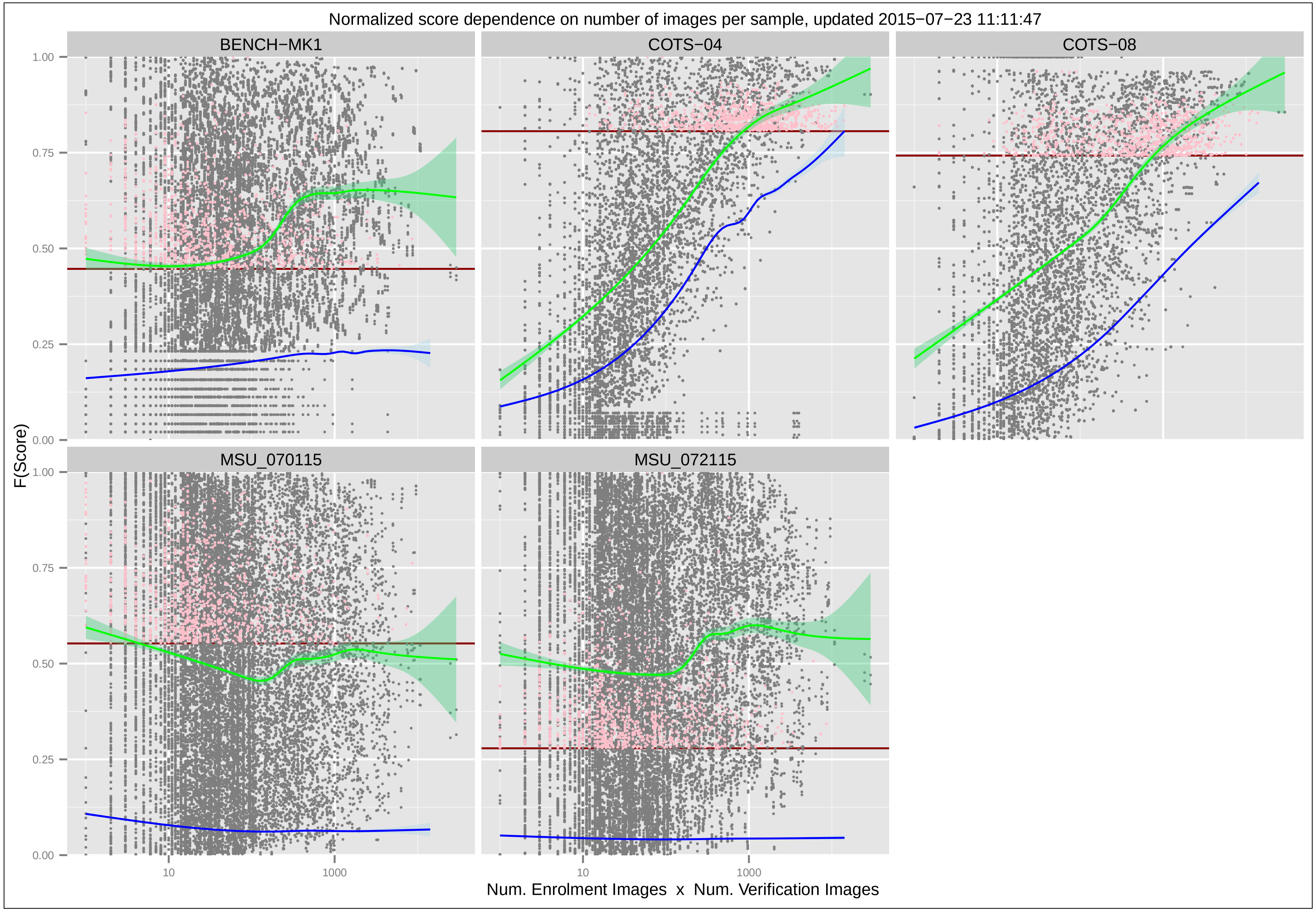

Figure 8: **Effect of multi-image samples:** Each panel plots normalized similarity scores against the product of the number of images available in the enrollment and verification samples. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. Normalized genuine scores are shown as grey dots. The green line shows a smoothed dependence of genuine score. The blue line shows the dependence for impostor scores. Only impostors above the red line, which corresponds to FNMR at the threshold which gives FMR = 0.01 globally, are plotted (in pink). Scores at or below zero are suppressed. The same trends exist if raw scores are plotted, and if the max(N_1, N_2) is used in place of $N_1 N_2$.

3.2 Effect of covariates

This section seeks to associate recognition outcome to properties of input images, or of people. Two variables are noted here. First and foremost, recognition algorithms are known to be intolerant of pose variation. Second, poor spatial resolution degrades recognition. In addition, fixed and time-varying subject-specific covariates such as facial hair, skin tone, age, and gender all have documented effects.

A necessary preface to this section. The IJB-A is not designed to produce a cause-and-effect model of recognition failure and, moreover, is inadequately sized. This section is included with the goal of revealing exceptional algorithm traits, such as pose invariance, and to assist developers. Notable algorithms should be submitted to NIST's larger-scale sequestered test.

3.2.1 Dependence on yaw angle

We adopt the yaw estimates from a government-owned algorithm as ground-truth. These are imperfect.

Given that a one-to-one verification score comes from the comparison of two samples, and that a sample generally includes multiple images, we first define methods of representing their joint pose.

- ▷ **Similar pose**: We compute the following statistic as being influential on matching.

  $$\theta = arg\, min f(x) \tag{4}$$

  where x is the tuple (i, j) referring to the i-th enrolment image, $1 \leq i \leq n_1$, and j-th verification image, $1 \leq i \leq n_2$, and the yaw-difference function is

  $$f(x) = |\theta_i - \theta_j| \tag{5}$$

  This definition is simply the smallest difference between any two images contained in the two input samples. So for if any pair of enrolment and verification images have the same pose, then $\theta = 0$. This statistic would clearly be relevant to the case where a naiive algorithm represents n_1 input images as n_1 separate feature vectors, and implements verification by comparison of $n_1 n_2$ feature vectors. Whether this statistic remains relevant to implementations that fuse information at the feature level is not clear.

  The effect of yaw difference on genuine scores is shown in Figure 9. The corresponding graphs for impostor scores appear in Figure 11.

- ▷ **Frontal pose**: An alternative model would be to assume that recognition is related to the presence of a frontal face. This is relevant because zero yaw faces are formally standardized as the canonical view, and the target of much development, and the default output view for morphable model approaches. We compute how different from frontal the two samples are:

  $$\theta = |\theta^{(1)} - \theta^{(2)}| \tag{6}$$

  where $\theta^{(1)}$ is the smallest (most frontal) yaw angle of the n_1 input images

  $$\theta^{(1)} = \min_i |\theta_i| \tag{7}$$

  and likewise for $\theta^{(2)}$.

  This method is appropriate in cases where a standard is reasonably expected to apply to one of the images. This is the case with operator-attended credential issuance for example, where $\theta^{(1)} \rightarrow 0$ is enforced via adjudication and recapture. Failure analysis then reduces to quality analysis when measurement of properties of one image is important and similarity score is held to then just depend on the verification sample i.e. $s \sim F(\theta^{(2)})$.

The effect of non-frontal pose on genuine scores is shown in Figure 10.

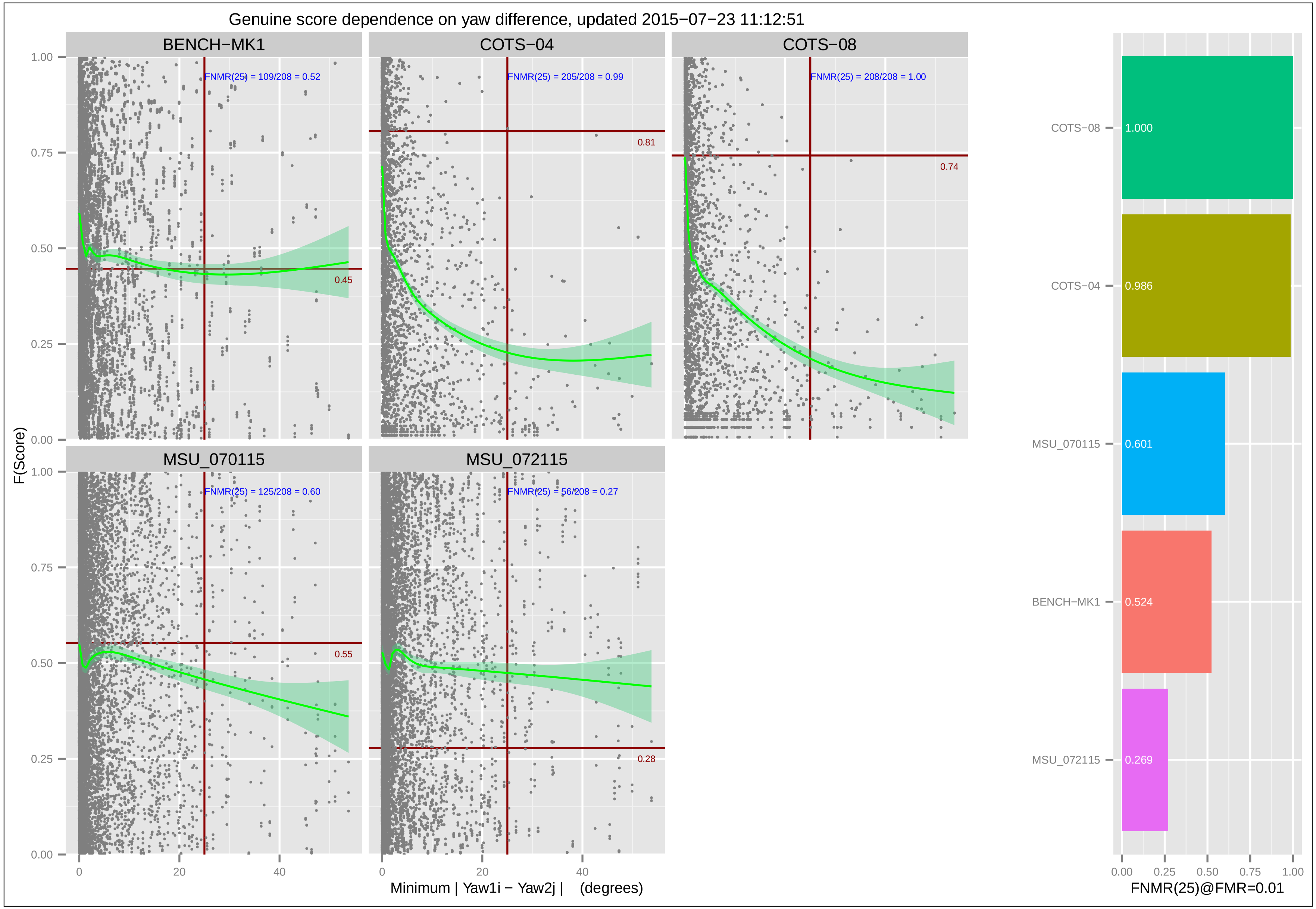

Figure 9: **Effect of yaw difference on genuine comparisons:** Each panel plots normalized similarity scores against the summary statistic given in equation 4. If the enrolment and verification samples both contain an image with the subject's head at say, -45 degrees, then this statistic is zero. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. Normalized scores are shown as grey dots. The green line shows a smoothed dependence of genuine score. The red line corresponds to FNMR at the threshold which gives FMR = 0.01 globally. Scores at or below zero are replaced with the global minimum score. The right panel and blue text show FNMR for comparisons whose mutual yaw difference is above 25 degrees. Some panels have points than others. This arises because across the 10 IJB-A splits some comparisons are repeated. Some algorithms give identical scores, some do not. Algorithms that train on each split give different scores for the same input template pair.

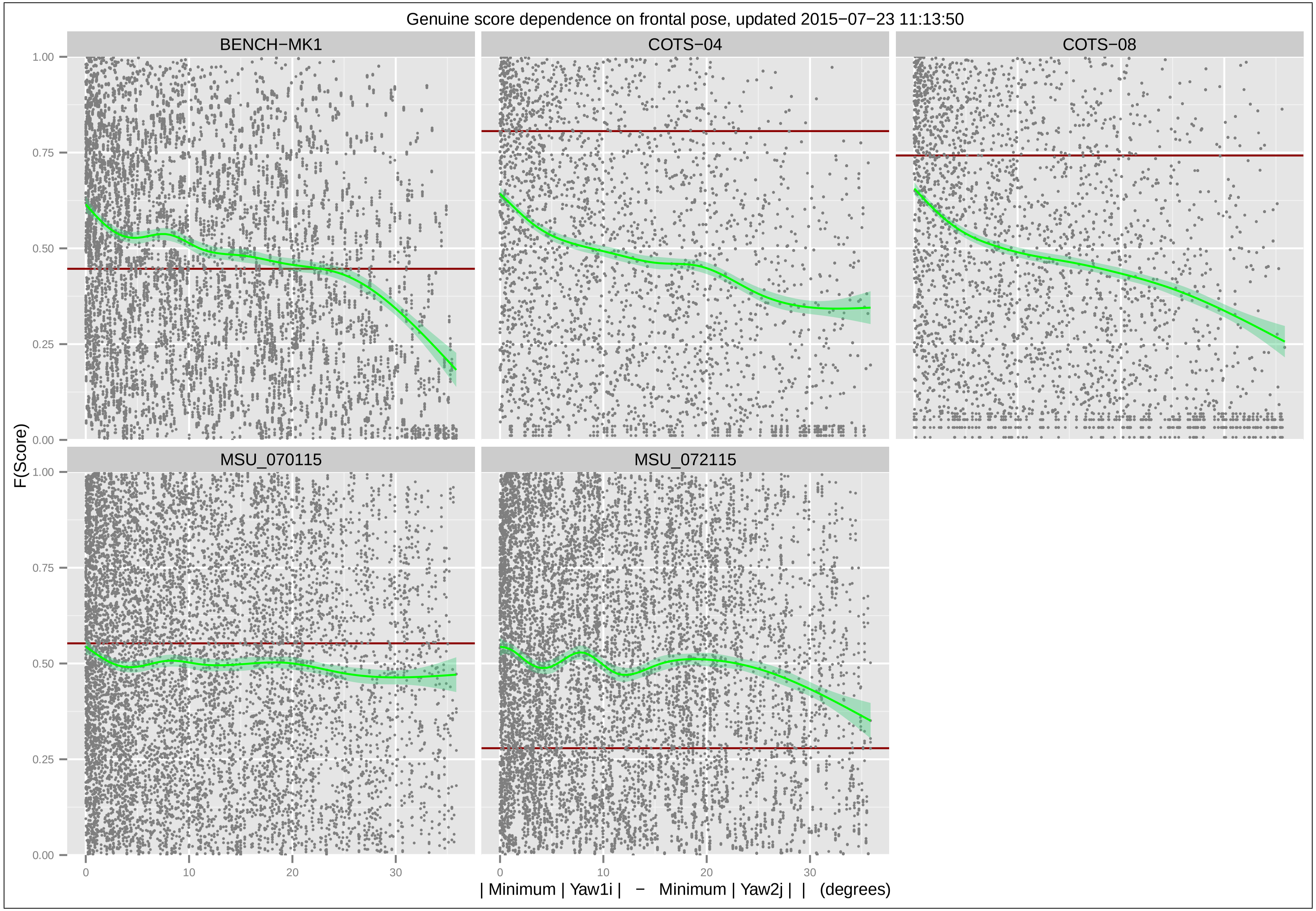

Figure 10: **Effect of non-frontal yaw on genuine comparisons:** Each panel plots normalized similarity scores against the summary statistic given in equation 6. The x-axis quantity is zero only if both the enrolment and verification samples contain images with zero yaw. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. Normalized genuine scores are shown as grey dots. The green line shows a smoothed dependence of genuine score. The red line corresponds to FNMR at the threshold which gives FMR = 0.01 globally. Scores at or below zero are replaced with the global minimum score. Some panels have points than others. This arises because across the 10 IJB-A splits some comparisons are repeated. Some algorithms give identical scores, some do not. Algorithms that train on each split give different scores for the same input template pair.

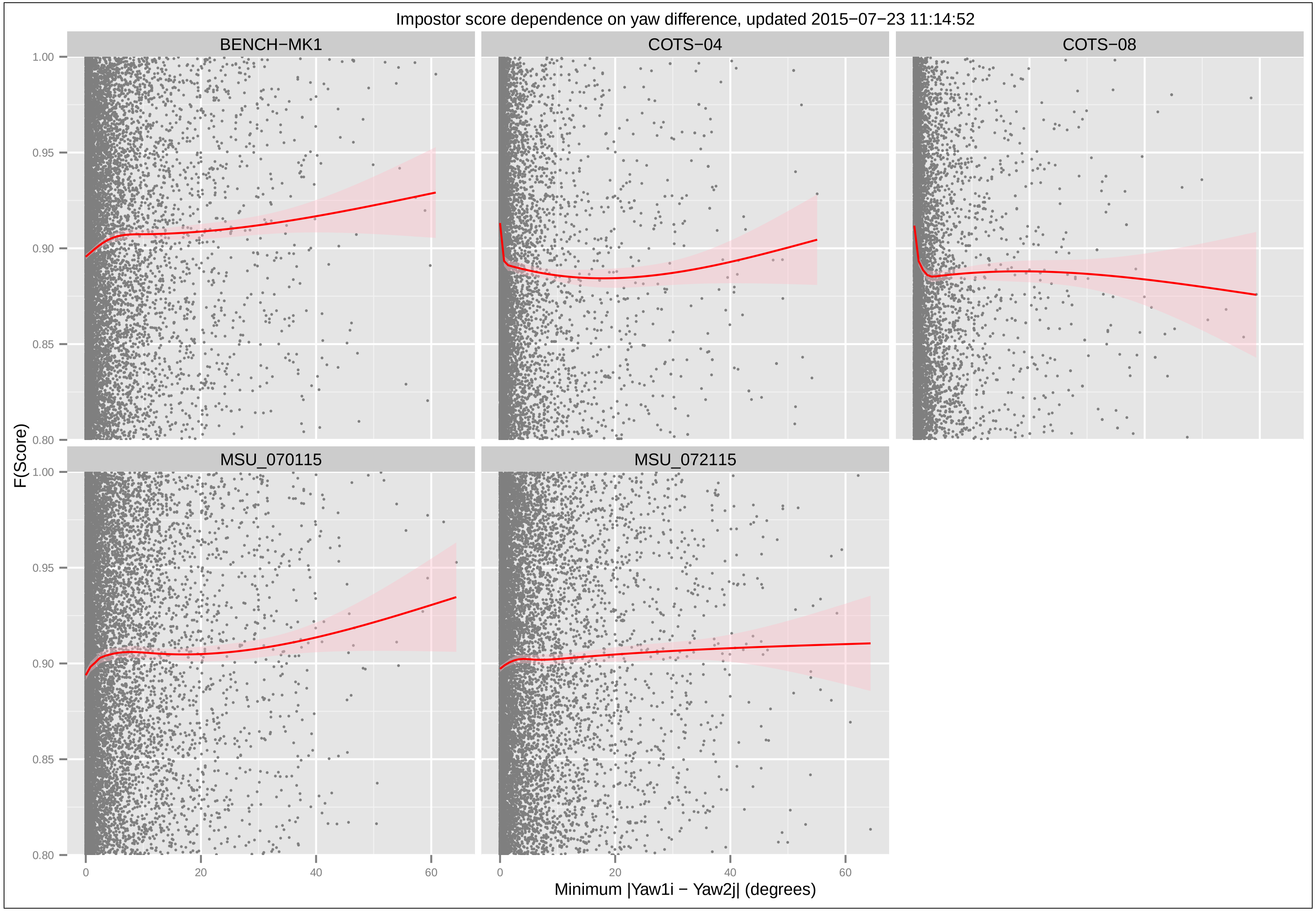

Figure 11: **Effect of yaw difference matching on impostor comparisons:** Each panel plots normalized impostor scores against the summary statistic given in equation 4. If the enrolment and verification samples both contain an image with the subject's head as say, -45 degrees, then this statistic is zero. The normalization is the probability integral transform (see sec. 2.4) computed over just impostor scores. Normalized scores are shown as grey dots. The red line shows a smoothed dependence of impostor score. Scores at or below zero are replaced with the minimum positive score.

3.2.2 Effect of resolution

Spatial resolution is known to have an effect on face recognition accuracy. However, optical resolution is usually unknown, and a weak proxy is used. This is the spatial sampling rate, usually stated as the interocular distance (IOD). In this report, given the disappearance of one eye due to adverse yaw angle, we have only a weaker proxy for resolution, the area of the bounding box. This information was manually produced.

Given n_i area measurements for sample $i = 1, 2$, we find the largest image area, on the (weak) assumption that this image will be of most utility to recognition:

$$A^{(i)} = \max \left\{ A_1^{(i)}, A_2^{(i)}, \ldots, A_{n_i}^{(i)} \right\} \tag{8}$$

We the consider that the smaller of the two areas

$$A = \min \left(A^{(1)},\, A^{(2)} \right) \tag{9}$$

will drive matching failure. This is a worst-of-the-best model. Finally, to ease interpretation we convert to a linear dimension via

$$D = \sqrt{A} \tag{10}$$

and actually plot that on a log scale via $D = \log_{10} \sqrt{1 + A}$.

The effect of spatial size is shown in Figure 12. Most algorithms give high genuine scores for larger faces. This is not true for at least one algorithm. Some algorithms exhbit a lower cutoff below which detection or recognition fails.

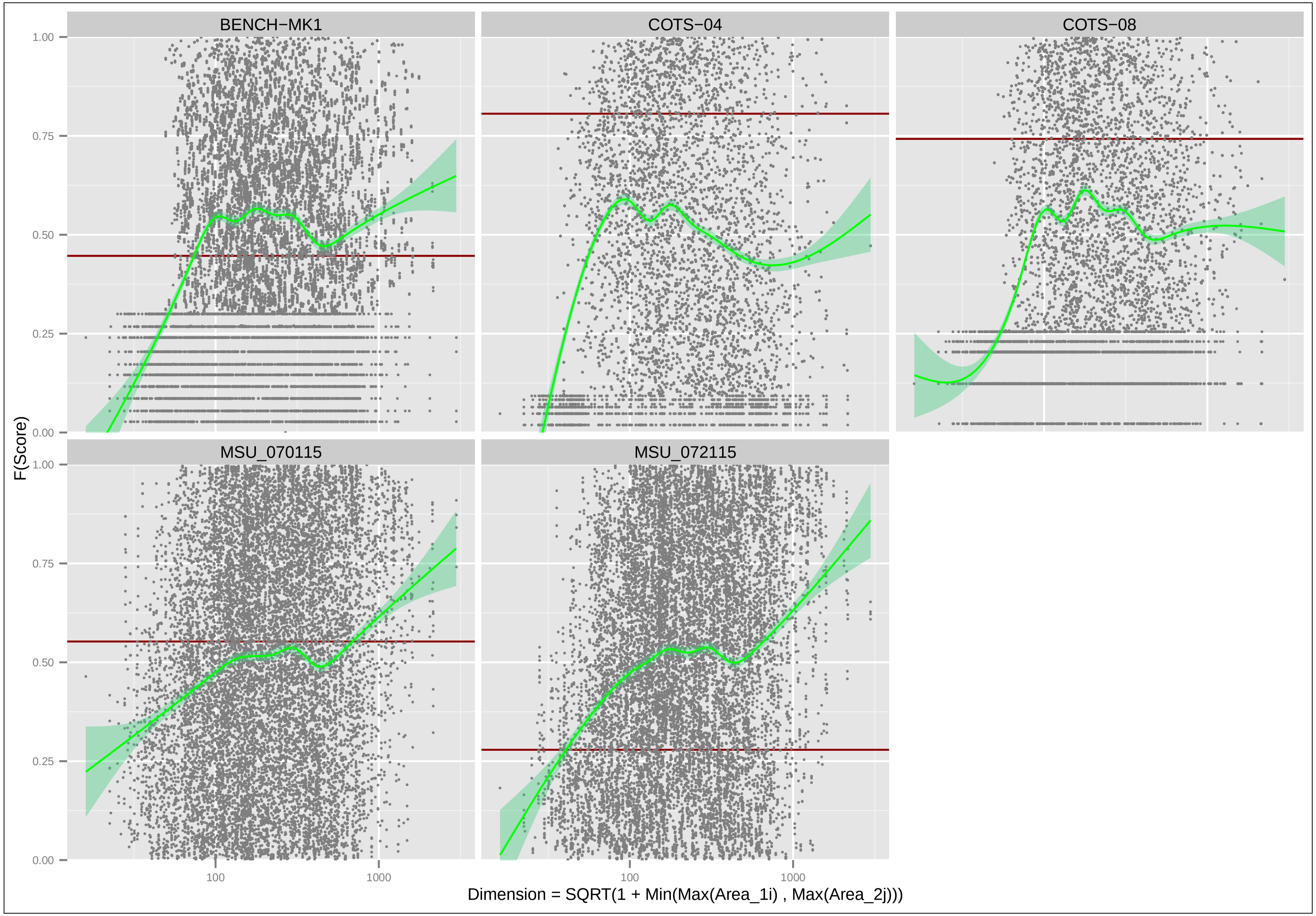

Figure 12: **Effect of bounding box area**: Each panel plots normalized genuine scores against the summary statistic given in equation 10. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. Normalized scores are shown as grey dots. The green line shows a smoothed dependence of genuine score. The red line corresponds to FNMR at the threshold which gives FMR = 0.01 globally. Scores at or below zero are replaced with the minimum positive score.

3.2.3 Effect of age

This section looks at accuracy by age group. It does not consider longitudinal ageing effects. Ages were assigned onto a six category scale. Each image was assessed by several observers.

Each sample is comprised of several images. In the IJB-A design these images are not necessarily contemporaneous. We associate the age annotations with similarity scores as follows. Given n_i age assessments for sample $i = 1, 2$, we find the most common age over all images.

The effect of age group is shown in the boxplots of Figure 13.

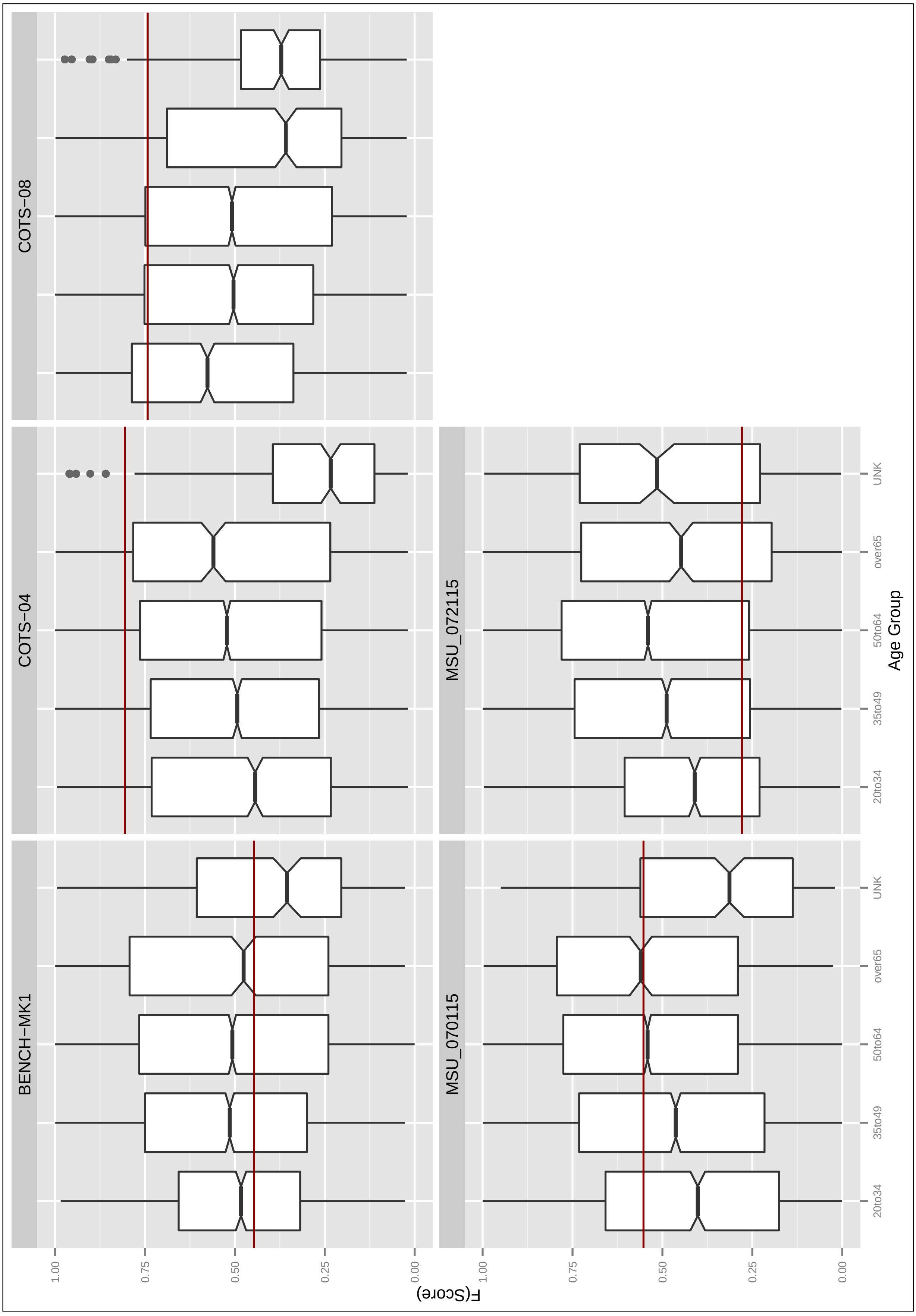

Figure 13: **Effect of skin tone**: Each panel plots normalized genuine scores against the most commonly assigned age group for the enrollment images. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. Non-overlapping boxplot notches are a statement of significant difference.

3.2.4 Effect of facial hair

Facial hair can be problematic as it occludes features and changes appearance.

We use manual annotations. The enrollment sample is assigned the label that is most often assigned to the n_1 constituent images. Likewise for the verification sample.

The effect of facial hair is shown in the boxplots of Figure 15. There is no clear indication that beards present a problem. However change does.

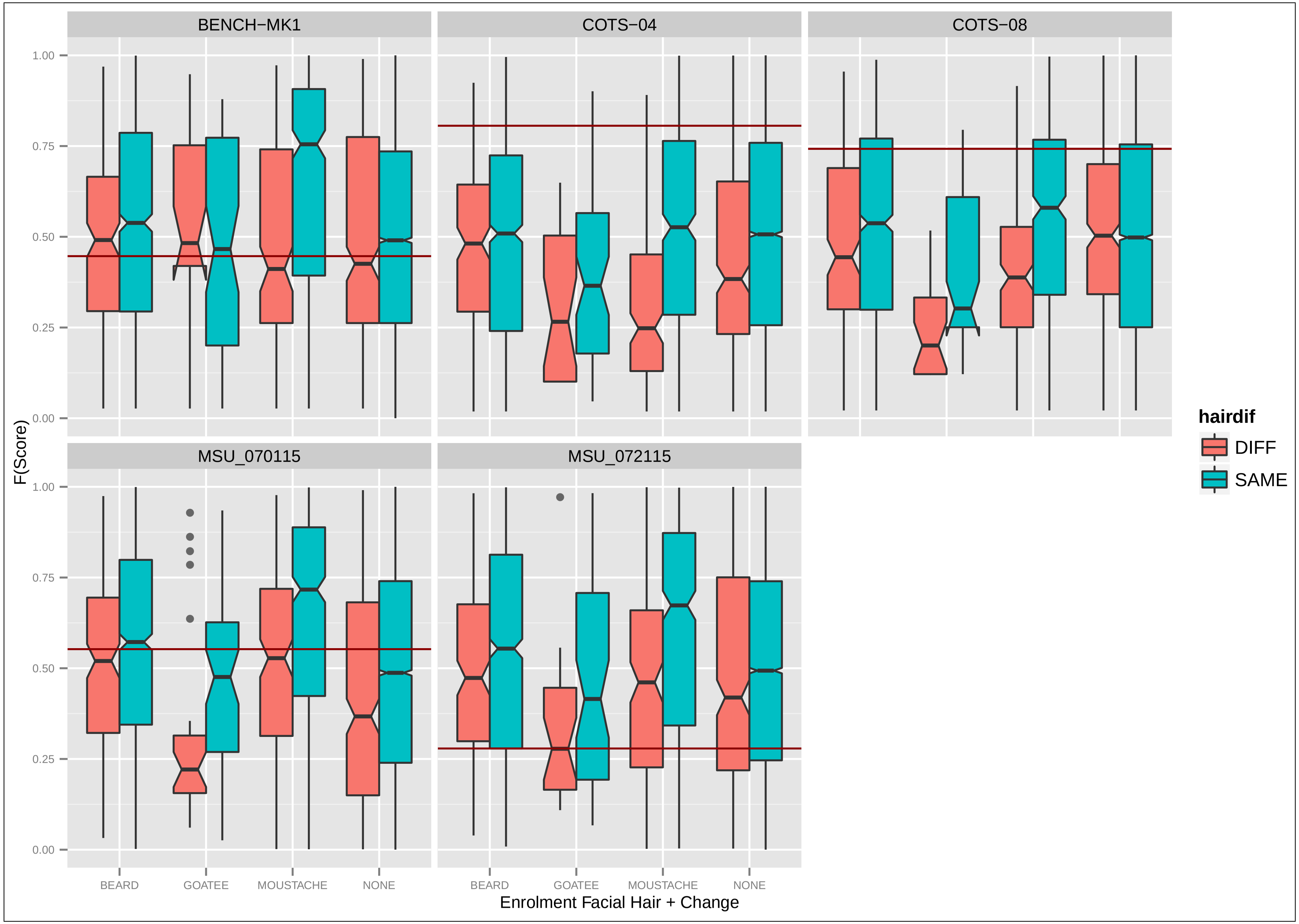

Figure 14: **Effect of facial hair**: Each panel gives boxplots of normalized genuine scores by facial hair type, and whether it changed between enrollment and verification. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. These values have some correlation with race. Non-overlapping boxplot notches are a statement of significant difference.

3.2.5 Effect of skintone

The color of the skin in images can affect face detection and landmark localization: White skin can be overexposed, and dark skin underexposed. This section gives results for manually-labelled skin tone assessments on a six-level scale going from light to dark.

We associate these skin tone values with similarity scores as follows. Given n_i skin tone assessments for sample $i = 1, 2$, we find the mean skin tone value over all images.

$$S^{(i)} = \frac{1}{n_i} \sum_{k=1}^{n_i} S_k^{(i)} \tag{11}$$

And then form the summary scalar

$$S = \frac{S^{(1)} + S^{(2)}}{2} \tag{12}$$

The effect of skin tone is shown in the boxplots of Figure 15. There is no consistent trend across algorithms.

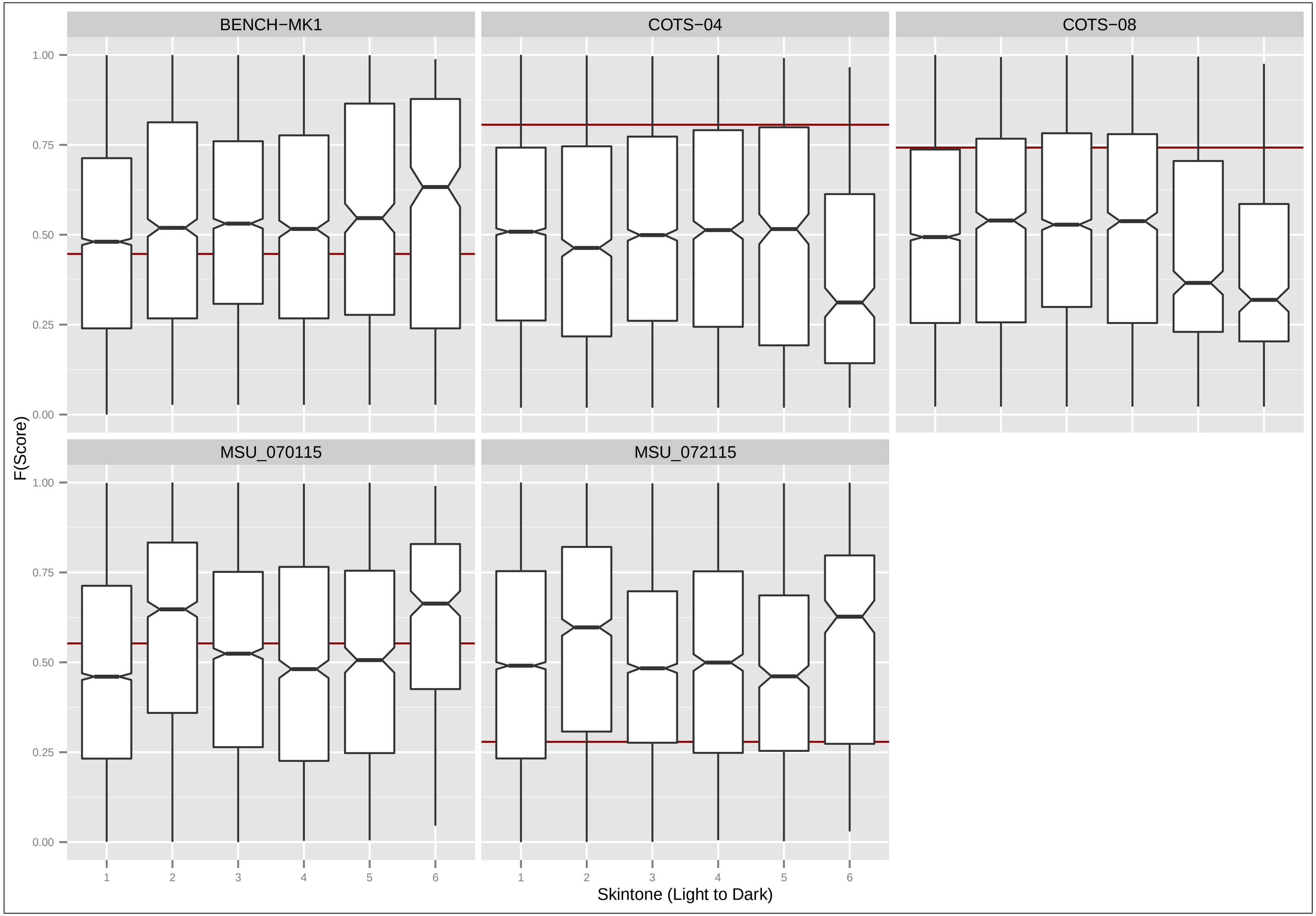

Figure 15: **Effect of skin tone**: Each panel plots normalized genuine scores against the aggregate skin tone means of equation 12. The normalization is the probability integral transform (see sec. 2.4) computed over just genuine scores. The skintone scale 1 . . . 6 goes from light to dark. These values have some correlation with race. Non-overlapping boxplot notches are a statement of significant difference. The IJB-A images are mostly correctly exposed.

4 References

[1] S. Curry, D. Founds, J. Marques, N. Orlans (Mitre), and C. Watson (NIST). Meds - multiple encounter deceased subject face database - nist special database 32. NIST Interagency Report 7679, National Institute of Standards and Technology, 2011. http://www.nist.gov/itl/iad/ig/sd32.cfm. 1

[2] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proc. IEEE CVPR*, June 2015. 1

[3] T. Mansfield. *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework*. JTC1/SC37/Working Group 5, August 2005. http://webstore.ansi.org. 2